

# Semi-supervised Deep Generative Modelling of Incomplete Multi-Modality Emotional Data

Changde Du

Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA  
University of CAS, Beijing, China  
duchangde@gmail.com

Changying Du, Hao Wang

360 Search Lab  
Beijing, China  
ducyatict@gmail.com  
cashenry@126.com

Jinpeng Li

Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA  
University of CAS, Beijing, China  
lijinpeng2015@ia.ac.cn

Wei-Long Zheng

Department of Computer Science and Engineering, SJTU  
Shanghai, China  
weilong@sjtu.edu.cn

Bao-Liang Lu

Department of Computer Science and Engineering, SJTU  
Shanghai, China  
bllu@sjtu.edu.cn

Huiguang He\*

Research Center for Brain-inspired Intelligence & National Laboratory of Pattern Recognition, CASIA  
huiguang.he@ia.ac.cn

## ABSTRACT

There are threefold challenges in emotion recognition. First, it is difficult to recognize human's emotional states only considering a single modality. Second, it is expensive to manually annotate the emotional data. Third, emotional data often suffers from missing modalities due to unforeseeable sensor malfunction or configuration issues. In this paper, we address all these problems under a novel multi-view deep generative framework. Specifically, we propose to model the statistical relationships of multi-modality emotional data using multiple modality-specific generative networks with a shared latent space. By imposing a Gaussian mixture assumption on the posterior approximation of the shared latent variables, our framework can learn the joint deep representation from multiple modalities and evaluate the importance of each modality simultaneously. To solve the labeled-data-scarcity problem, we extend our multi-view model to semi-supervised learning scenario by casting the semi-supervised classification problem as a specialized missing data imputation task. To address the missing-modality problem, we further extend our semi-supervised multi-view model to deal with incomplete data, where a missing view is treated as a latent variable and integrated out during inference. This way, the proposed overall framework can utilize all available (both labeled and unlabeled, as well as both complete and incomplete) data to improve its generalization ability. The experiments conducted on two real multi-modal emotion datasets demonstrated the superiority of our framework.

\*Dr. Huiguang He is the corresponding author and he is also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences and the University of Chinese Academy of Sciences.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*MM'18*, Seoul, Korea

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5665-7/18/10...\$15.00

<https://doi.org/10.1145/3240508.3240528>

## KEYWORDS

Multi-view semi-supervised learning; deep generative model; incomplete data; multi-modal emotion recognition

### ACM Reference Format:

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised Deep Generative Modelling of Incomplete Multi-Modality Emotional Data. In *2018 ACM Multimedia Conference (MM'18)*, October, 2018, Seoul, Korea. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3240508.3240528>

## 1 INTRODUCTION

With the development of human-computer interaction (HCI), emotion recognition has become increasingly important. Since human's emotion contains many nonverbal cues, various modalities ranging from facial expressions, body gesture, voice to physiological signals can be used as the indicators of emotional states [5, 24]. In real-world applications, it is difficult to recognize human's emotional states only considering a single modality, because signals from different modalities represent different aspects of emotion and provide complementary information. Recent studies show that integrating multiple modalities can significantly boost the emotion recognition accuracy [18, 26, 31]. The most successful approach to fuse the information from multiple modalities is based on deep multi-view representation learning [21, 33, 37]. E.g., [23] proposed a joint density model for emotion analysis with a multi-modal deep Boltzmann machine (DBM) [33]. This multi-modal DBM is exploited to model the joint distribution over visual, auditory, and textual features. [17] proposed a multi-modal emotion recognition method by using multi-modal autoencoders (MAE) [21], in which the joint representations of Electroencephalogram (EEG) and eye movement signals were extracted. Nevertheless, there are still limitations with these deep multi-modal emotion recognition methods, e.g., their performances depend on the amount of labeled data and they could not handle incomplete data.

By using the modern sensing equipments, we can easily collect massive emotion-related data from multiple modalities. But, the data labeling procedure requires lots of manual efforts. Therefore, in most cases only a small set of labeled samples is available, while

the majority of whole dataset is left unlabeled. In addition to challenges with insufficient labeled data, one must often address the incomplete-data problem, i.e., not all modalities are available for every data point. Generally, we can identify various causes for incomplete data. E.g., unforeseeable sensor malfunction may fail to collect sensing information, thus providing us incomplete data with one or more missing modalities. Traditional multi-modal emotion recognition approaches [17, 18, 23] only utilized the limited amount of labeled data, which may result in severe overfitting. Also, most of them neglect the missing modality issue, which greatly limits their applications in real-world scenarios. The most attractive way to deal with the aforementioned issues is semi-supervised learning (SSL) with incomplete data. SSL can improve model's generalization ability by exploiting both labeled and unlabeled data simultaneously [10, 28, 43], and learning from incomplete data can guarantee the robustness of the emotion recognition system [35].

In this paper, we show that the problems mentioned above can be resolved under a unified multi-view deep generative framework. For modeling the statistical relationships of multi-modality emotional data, a shared latent variable is transformed by different modality-specific generative networks to different data views (modalities). Instead of treating each view equally, we impose a non-uniformly weighted Gaussian mixture assumption on the posterior approximation of the shared latent variables. This is critical for inferring the joint latent representation and the weight factor of each view from multiple modalities. During optimization, a second lower bound to the variational lower bound is derived to address the intractable entropy of a mixed Gaussians. To leverage the contextual information in the unlabeled data to augment the limited labeled data, we then extend our multi-view framework to SSL scenario. It is achieved by casting the semi-supervised classification problem as a specialized missing data imputation task. Specifically, we treat the unknown labels as latent variables and estimate them within a multi-view auto-encoding variational Bayes framework. We further extend the proposed SSL algorithm to the incomplete-data case by introducing latent variables for the missing views. Besides the unknown labels, the missing views are also integrated out so that the marginal likelihood is maximized with respect to model parameters. In this way, our SSL algorithm can utilize all available data: both labeled and unlabeled, as well as both complete and incomplete. Since the category information and the uncertainty of missing view are taken into account in the training process, our SSL algorithm is more powerful than traditional missing view imputation methods [6, 8, 25, 37]. We finally demonstrate the superiority of our framework and provide insightful observations on two real multi-modal emotion datasets.

## 2 RELATED WORK

Multi-modal approaches have been widely implemented for emotion recognition [17, 18, 23, 26, 31, 34, 44]. E.g., [26] used a multi-modal deep belief network (DBN) to extract features from face, body gesture, voice and physiological signals for emotion classification. [18] classified the combination of EEG and eye movement signals into three affective states. But, very few of them explored SSL. To the best of our knowledge, only [43] proposed an enhanced multi-modal co-training algorithm for semi-supervised emotion

recognition, but its shallow structure is hard to capture the high-level correlation between different modalities. In addition, most prior work in this field assumes that all modalities are available at all times [43, 44], which is not realistic in practical environments. In contrast to the above methods, our framework naturally allows us to perform multi-modal emotion recognition within SSL and incomplete-data situations.

The variational autoencoder (VAE) [15, 27] is one of the most popular deep generative models (DGMs). VAE has shown great advantages in semi-supervised classification [13, 20]. E.g., Kingma et al. [13] proposed a semi-supervised VAE (M2) by modeling the joint distribution over data and labels. Maaløe et al. proposed the auxiliary DGMs (ADGM and SDGM) [20] by introducing auxiliary variables, which improve the variational approximation. However, these models cannot effectively deal with multi-view data, especially in incomplete-view case. Our proposed semi-supervised multi-view DGMs distinguish our method from all existing ones using VAE framework [3, 14, 20, 29, 38].

Incomplete-data problem is often circumvented via imputation methods [1, 36, 39, 40, 42]. Common imputation schemes include matrix completion [8, 9, 11] and autoencoder-based methods [6, 19, 30, 37]. Matrix completion methods, such as SoftImputeALS [8], focus on imputing the missing entries of a partially observed matrix based on assumption that the completed matrix has a low-rank structure. Matrix completion methods often assume data is missing at random (MAR), which might not be optimal for our problem where modalities are missing at continuous blocks. On the other hand, autoencoder-based methods, such as DCCAE [37] and CorrNet [6], exploit the connections between views, enabling the incomplete view to be restored with the help of the complete view. Besides low-rank structure of the data matrix and the connections between views, category information is also important for missing view imputation tasks, though category labels may be partially observed. So far, very few algorithms [25, 41] can estimate the missing view under the SSL scenario. Although CoNet [25] utilized deep neural networks (DNNs) to predict the missing view based on existing views and partially observed labels, its feedforward structure could not integrate multiple views effectively in classification. Additionally, most previous works treat the missing data as fixed values and hence ignore the uncertainty of the missing data. Unlike them, our SiMVAE essentially performs infinite imputations by integrating out the missing data.

## 3 METHODOLOGY

In this section, we first develop a multi-view variational autoencoder (MVAE) model for fusing multi-modality emotional data. Based on MVAE, we further build a semi-supervised emotion recognition algorithm. Finally, we develop a more robust semi-supervised algorithm to address the incomplete multi-modality emotional data. For simplicity we restrict further discussion to the case of two views, though all the proposed methods can be extended to more than two views. Assume we are faced with multi-view data that appears as pairs  $(\mathbf{x}, y) = (\{\mathbf{x}^{(v)}\}_{v=1}^2, y)$ , with observation  $\mathbf{x}^{(v)}$  from the  $v$ -th view and the corresponding class label  $y$ .

### 3.1 Multi-view Variational Autoencoder

**3.1.1 DNN-parameterized Likelihoods.** We assume that multiple data views (modalities)  $\{\mathbf{x}^{(v)}\}_{v=1}^2$  are generated independently from a shared latent space with multiple view-specific generative networks. Specifically, we assume a shared latent variable  $\mathbf{z}$  generates  $\mathbf{x}^{(v)}$  with the following generative model  $P_1$  (cf. Figure 1a):

$$p_{\theta^{(v)}}(\mathbf{x}^{(v)}|\mathbf{z}) = f(\mathbf{x}^{(v)}; \mathbf{z}, \theta^{(v)}), \quad v \in \{1, 2\}, \quad (1)$$

where  $f(\mathbf{x}^{(v)}; \mathbf{z}, \theta^{(v)})$  is a suitable likelihood function (e.g. a Gaussian for continuous observation or Bernoulli for binary observation), which is formed by a non-linear transformation of the latent variable  $\mathbf{z}$ . This non-linear transformation is essential to allow for higher moments of the data to be captured by the density model, and we choose these non-linear functions to be DNNs, referred to as the generative networks, with parameters  $\{\theta^{(v)}\}_{v=1}^2$ . Note that, the likelihoods for different views are assumed to be independent of each other, with potentially different DNN types for different modalities.

**3.1.2 Gaussian Prior and Gaussian Mixture Posterior.** In vanilla VAE [15, 27], which can only handle single-view data, both the prior  $p(\mathbf{z})$  and the approximate posterior  $q_{\phi}(\mathbf{z}|\mathfrak{X})$  are assumed to be Gaussian distributions in order to maintain mathematical and computational tractability. Although this assumption has led to favorable results on several tasks, it is clearly a restrictive and often unrealistic assumption. Specifically, the choice of a Gaussian distribution for  $p(\mathbf{z})$  and  $q_{\phi}(\mathbf{z}|\mathfrak{X})$  imposes a strong uni-modal structure assumption on the latent space. However, for data distributions that are strongly multi-modal, the uni-modal Gaussian assumption inhibits the model's ability to extract and represent important structure in the data. To improve the flexibility of the model, one way is to impose a Mixture of Gaussians (MoG) assumption on  $p(\mathbf{z})$ . However, it has the risk of creating separate "islands" of discontinuous manifolds that may break the meaningfulness of the representation in the latent space.

To learn more powerful and expressive models (in particular, models with multi-modal latent variable structures for multi-modal emotion recognition applications) we seek a MoG for  $q_{\phi}(\mathbf{z}|\mathfrak{X})$ , while preserving  $p(\mathbf{z})$  as a standard Gaussian. Thus the prior distribution and the inference model  $Q_1$  (cf. Figure 1b) are defined as:  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ ,

$$q_{\phi}(\mathbf{z}|\mathfrak{X}) = \sum_{v=1}^2 \lambda^{(v)} \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}), \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)})\right), \quad (2)$$

where the mean  $\boldsymbol{\mu}_{\phi^{(v)}}$  and the covariance  $\boldsymbol{\Sigma}_{\phi^{(v)}}$  are nonlinear functions of the observation  $\mathbf{x}^{(v)}$ , with variational parameter  $\phi^{(v)}$ . As in our generative model, we choose these nonlinear functions to be DNNs, referred to as the inference networks.  $\lambda^{(v)}$  is the non-negative normalized weight factor for the  $v$ -th view, i.e.,  $\lambda^{(v)} > 0$  and  $\sum_{v=1}^2 \lambda^{(v)} = 1$ . Note that, Gershman et al. [7] proposed a non-parametric variational inference method by simply assuming the variational distribution to be a uniformly weighted Gaussian mixture. However, treating each component equally will lose flexibility in fusing multiple data views. Instead of treating each view equally, our non-uniformly weighted Gaussian mixture assumption can

weight each view automatically in subsequent emotion recognition tasks, which is useful to identify the importance of each view.

### 3.2 Semi-supervised Multi-modal Emotion Recognition

Although many supervised emotion recognition algorithms exist (see [24] for a thorough literature review), very few semi-supervised algorithms have been proposed to improve the recognition performance by utilizing both labeled and unlabeled data. Here we extend MVAE by introducing a conditional probabilistic distribution for the unknown labels to obtain a semi-supervised multi-view classification algorithm.

**3.2.1 Generative model  $P_2$ .** Since the emotional data is continuous, we choose the Gaussian likelihoods. Then our generative model  $P_2$  (cf. Figure 1c) is defined as  $p(y)p(\mathbf{z})\prod_{v=1}^2 p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})$ :

$$p(y) = \text{Cat}(y|\boldsymbol{\pi}), \quad p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}), \quad (3)$$

$$p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_{\theta^{(v)}}(y, \mathbf{z}), \text{diag}(\boldsymbol{\sigma}_{\theta^{(v)}}^2(y, \mathbf{z}))\right),$$

where  $\text{Cat}(\cdot)$  denotes the categorical distribution,  $y$  is treated as a latent variable for the unlabeled data points, and the mean  $\boldsymbol{\mu}_{\theta^{(v)}}$  and variance  $\boldsymbol{\sigma}_{\theta^{(v)}}^2$  are nonlinear functions of  $y$  and  $\mathbf{z}$ , with parameter  $\theta^{(v)}$ .

**3.2.2 Inference model  $Q_2$ .** The inference model  $Q_2$  (cf. Figure 1d) is defined as  $q_{\phi}(y|\mathfrak{X})q_{\phi}(\mathbf{z}|\mathfrak{X}, y)$ :

$$q_{\phi}(y|\mathfrak{X}) = \text{Cat}(y|\boldsymbol{\pi}_{\phi}(\mathfrak{X})), \quad (4)$$

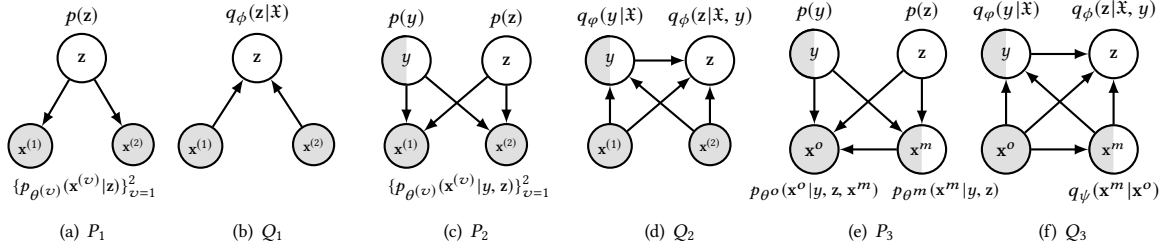
$$q_{\phi}(\mathbf{z}|\mathfrak{X}, y) = \sum_{v=1}^2 \lambda^{(v)} \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}, y), \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)}, y)\right),$$

where  $q_{\phi}(y|\mathfrak{X})$  is the introduced conditional distribution for  $y$ , and  $q_{\phi}(\mathbf{z}|\mathfrak{X}, y)$  is assumed to be a mixture of Gaussians to combine the information from multiple data views and the label. Intuitively,  $q_{\phi}(\mathbf{z}|\mathfrak{X}, y)$ ,  $p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})$  and  $q_{\phi}(y|\mathfrak{X})$  correspond to the encoder, decoder and classifier, respectively. For brevity, we omit the explicit dependencies on  $\mathbf{x}^{(v)}$ ,  $y$  and  $\mathbf{z}$  for the moment variables mentioned above hereafter. In principle,  $\boldsymbol{\mu}_{\theta^{(v)}}$ ,  $\boldsymbol{\sigma}_{\theta^{(v)}}^2$ ,  $\boldsymbol{\pi}_{\phi}$ ,  $\boldsymbol{\mu}_{\phi^{(v)}}$  and  $\boldsymbol{\Sigma}_{\phi^{(v)}}$  can be implemented by various DNN models, e.g., multi-layer perceptrons and convolutional neural networks.

**3.2.3 Objective function.** In the semi-supervised setting, there are two lower bounds for the *labeled* and *unlabeled* cases, respectively. The variational lower bound on the marginal likelihood for a single *labeled* data point is

$$\begin{aligned} \log p_{\theta}(\mathfrak{X}, y) &\geq \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathfrak{X}, y)} \left[ \log \frac{p_{\theta}(\mathfrak{X}, y, \mathbf{z})}{q_{\phi}(\mathbf{z}|\mathfrak{X}, y)} \right] \\ &\geq \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathfrak{X}, y)} \left[ \sum_{v=1}^2 \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}) + \log p(y) \right. \\ &\quad \left. + \log p(\mathbf{z}) \right] - \sum_{v=1}^2 \lambda^{(v)} \cdot \log \left( \sum_{l=1}^2 \lambda^{(l)} \cdot \omega_{v,l} \right) \\ &\equiv -\mathcal{L}(\mathfrak{X}, y), \end{aligned} \quad (5)$$

where  $\omega_{v,l} = \mathcal{N}(\boldsymbol{\mu}_{\phi^{(v)}}|\boldsymbol{\mu}_{\phi^{(l)}}, \boldsymbol{\Sigma}_{\phi^{(v)}} + \boldsymbol{\Sigma}_{\phi^{(l)}})$ . It should be noted that, the Shannon entropy  $\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathfrak{X}, y)}[-\log q_{\phi}(\mathbf{z}|\mathfrak{X}, y)]$  is hard to



**Figure 1: Graphical models of the proposed algorithms: (a, b) multi-view variational autoencoder (MVAE); (c, d) semi-supervised MVAE (SMVAE); (e, f) semi-supervised incomplete MVAE (SiMVAE). In (e) and (f), we partition the two-view data point (i.e.,  $\tilde{\mathbf{x}} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\}$ ) into an observed view  $\mathbf{x}^o$  and a missing view  $\mathbf{x}^m$  (i.e.,  $\tilde{\mathbf{x}} = \{\mathbf{x}^o, \mathbf{x}^m\}$ ). Both  $y$  and  $\mathbf{x}^m$  are partially observed.**

compute analytically, and we have used the Jensen's inequality to derive a lower bound of it (see Supplementary Material Section A for details). For *unlabeled* data point, the variational lower bound on the marginal likelihood can be given by:

$$\begin{aligned} \log p_{\theta}(\tilde{\mathbf{x}}) &\geq \mathbb{E}_{q_{\phi}(y, z | \tilde{\mathbf{x}})} \left[ \log \frac{p_{\theta}(\tilde{\mathbf{x}}, y, z)}{q_{\phi}(y, z | \tilde{\mathbf{x}})} \right] \\ &= \mathbb{E}_{q_{\phi}(y | \tilde{\mathbf{x}})} \left[ -\mathcal{L}(\tilde{\mathbf{x}}, y) - \log q_{\phi}(y | \tilde{\mathbf{x}}) \right] \equiv -\mathcal{U}(\tilde{\mathbf{x}}), \end{aligned} \quad (6)$$

with  $q_{\phi}(y, z | \tilde{\mathbf{x}}) = q_{\phi}(y | \tilde{\mathbf{x}})q_{\phi}(z | \tilde{\mathbf{x}}, y)$ .

Therefore, the objective function for the entire dataset is:

$$\mathcal{J}_{\text{SMVAE}} = \underbrace{\sum_{(\tilde{\mathbf{x}}, y) \in S_l} \mathcal{L}(\tilde{\mathbf{x}}, y)}_{\text{labeled}} + \underbrace{\sum_{\tilde{\mathbf{x}} \in S_u} \mathcal{U}(\tilde{\mathbf{x}})}_{\text{unlabeled}}, \quad (7)$$

where  $S_l$  and  $S_u$  denote *labeled* and *unlabeled* dataset, respectively. The classification accuracy can be improved by introducing an explicit classification loss for labeled data, and the extended objective function is now:

$$\mathcal{F}_{\text{SMVAE}} = \mathcal{J}_{\text{SMVAE}} + \alpha \cdot \sum_{(\tilde{\mathbf{x}}, y) \in S_l} \left[ -\log q_{\phi}(y | \tilde{\mathbf{x}}) \right], \quad (8)$$

where  $\alpha$  is a weight parameter between generative and discriminative learning. We set  $\alpha = c \cdot \frac{(N_l + N_u)}{N_l}$ , where  $c$  is a scaling constant, and  $N_l$  and  $N_u$  are the numbers of labeled and unlabeled data points in one minibatch, respectively. Note that, the classifier  $q_{\phi}(y | \tilde{\mathbf{x}})$  is also used at test phase for the prediction of unseen data. Eq. (8) provides a unified objective function for optimizing the parameters of encoder, decoder and classifier networks.

**3.2.4 Parameter optimization.** Parameter optimization can be done jointly by using the stochastic backpropagation technique [15, 27]. The reparameterization trick [13, 15] is a vital component, because it allows us to take derivative of  $\mathbb{E}_{q_{\phi}(z | \tilde{\mathbf{x}}, y)} [\log p_{\theta^{(v)}}(\mathbf{x}^{(v)} | y, z)]$  w.r.t. the variational parameters  $\phi$ . However, the use of Gaussian mixture for variational posterior distribution  $q_{\phi}(z | \tilde{\mathbf{x}}, y)$  makes it infeasible to apply the reparameterization trick directly. It can be shown that, for any  $v \in \{1, 2\}$ ,  $\mathbb{E}_{q_{\phi}(z | \tilde{\mathbf{x}}, y)} [\log p_{\theta^{(v)}}(\mathbf{x}^{(v)} | y, z)]$  can

be rewritten, using the location-scale transformation for the Gaussian distribution, as:

$$\begin{aligned} &\mathbb{E}_{q_{\phi}(z | \tilde{\mathbf{x}}, y)} [\log p_{\theta^{(v)}}(\mathbf{x}^{(v)} | y, z)] \\ &= \sum_{l=1}^2 \lambda^{(l)} \mathbb{E}_{\mathcal{N}(\epsilon^{(l)} | 0, \mathbf{I})} \left[ \log p_{\theta^{(v)}}(\mathbf{x}^{(v)} | y, \boldsymbol{\mu}_{\phi^{(l)}} + \mathbf{R}_{\phi^{(l)}} \epsilon^{(l)}) \right], \end{aligned} \quad (9)$$

where  $\mathbf{R}_{\phi^{(l)}} \mathbf{R}_{\phi^{(l)}}^{\top} = \Sigma_{\phi^{(l)}}$  and  $l \in \{1, 2\}$ . While the expectations on the right hand side still cannot be solved analytically, their gradients w.r.t.  $\theta^{(v)}$ ,  $\phi^{(l)}$  and  $\lambda^{(l)}$  can be efficiently estimated using Monte-Carlo method (see Supplementary Material Section B for details). The gradients of the objective function (Eq. (8)) can then be computed by using the chain rule and the derived Monte-Carlo estimators.

### 3.3 Handling Incomplete Data

In the above discussion it is assumed that all modalities are available for every data point. In practice, however, many samples generally have incomplete modalities (i.e., with one or more missing modalities) [35]. In light of this, we further develop a semi-supervised incomplete multi-view classification algorithm (SiMVAE). For simplicity, we assume only one view (either  $\mathbf{x}^{(1)}$  or  $\mathbf{x}^{(2)}$ ) is incomplete, though our model can be easily extended to more sophisticated cases. We partition each data point into an observed view  $\mathbf{x}^o$  and a missing view  $\mathbf{x}^m$  (i.e.,  $\tilde{\mathbf{x}} = \{\mathbf{x}^o, \mathbf{x}^m\}$ ).

**3.3.1 Generative model  $P_3$ .** In this setting, only a subset of the samples have complete views and corresponding labels. We regard both the unknown label  $y$  and the missing view  $\mathbf{x}^m$  as latent variables. Then our generative model  $P_3$  (cf. Figure 1e) is defined as  $p(y)p(z)p_{\theta^m}(\mathbf{x}^m | y, z)p_{\theta^o}(\mathbf{x}^o | y, z, \mathbf{x}^m)$ :

$$\begin{aligned} p_{\theta^m}(\mathbf{x}^m | y, z) &= \mathcal{N} \left( \boldsymbol{\mu}_{\theta^m}(y, z), \text{diag}(\sigma_{\theta^m}^2(y, z)) \right), \\ p_{\theta^o}(\mathbf{x}^o | y, z, \mathbf{x}^m) &= \mathcal{N} \left( \boldsymbol{\mu}_{\theta^o}(y, z, \mathbf{x}^m), \text{diag}(\sigma_{\theta^o}^2(y, z, \mathbf{x}^m)) \right), \end{aligned} \quad (10)$$

where  $p_{\theta^m}(\cdot)$  and  $p_{\theta^o}(\cdot)$  are DNNs with parameters  $\theta^m$  and  $\theta^o$ , respectively.  $p(y)$  and  $p(z)$  are defined as in Eq. (3).

**3.3.2 Inference model  $Q_3$ .** As multi-modality emotional data are collected from the same subject, there must be some underlying relationships between modalities, though they focus on different information. Given the observed modality, the estimation of missing

modality is feasible if we capture the relationships between modalities. Therefore, the inference model  $Q_3$  (cf. Figure 1f) is defined as  $q_\psi(\mathbf{x}^m|\mathbf{x}^o)q_\varphi(y|\mathfrak{X})q_\phi(\mathbf{z}|\mathfrak{X}, y)$ , with

$$q_\psi(\mathbf{x}^m|\mathbf{x}^o) = \mathcal{N}\left(\boldsymbol{\mu}_\psi(\mathbf{x}^o), \text{diag}(\boldsymbol{\sigma}_\psi^2(\mathbf{x}^o))\right), \quad (11)$$

where  $q_\psi(\cdot)$  is a DNN with parameter  $\psi$ .  $q_\varphi(y|\mathfrak{X})$  and  $q_\phi(\mathbf{z}|\mathfrak{X}, y)$  are defined as in Eq. (4). Intuitively, we formulate the missing view imputation as a conditional distribution estimation task (conditioned on the observed view). Compared with existing single imputation methods [6, 19, 30], our model essentially performs infinite imputations and hence takes the uncertainty of the missing data into account. To obtain a single imputation of  $\mathbf{x}^m$  rather than the full conditional distribution one can evaluate  $\mathbf{x}^m = \mathbb{E}[q_\psi(\mathbf{x}^m|\mathbf{x}^o)]$ .

**3.3.3 Objective function.** In semi-supervised incomplete multi-view setting, there are four lower bounds for the *labeled-complete*, *labeled-incomplete*, *unlabeled-complete* and *unlabeled-incomplete* cases, respectively.

Similar to Eq. (5), the variational lower bound on the marginal likelihood for a single *labeled-complete* data point is

$$\begin{aligned} \log p_\theta(\mathfrak{X}, y) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathfrak{X}, y)}[\log p_\theta(\mathbf{x}^o|\mathbf{x}^m, y, \mathbf{z}) + \log p(y) \\ &+ \log p_\theta(\mathbf{x}^m|y, \mathbf{z}) + \log p(\mathbf{z})] - \sum_{v=1}^2 \lambda^{(v)} \cdot \log\left(\sum_{l=1}^2 \lambda^{(l)} \cdot \omega_{v,l}\right) \\ &\equiv -\mathcal{LC}(\mathfrak{X}, y), \end{aligned} \quad (12)$$

where  $\omega_{v,l} = \mathcal{N}(\boldsymbol{\mu}_{\phi^{(v)}}|\boldsymbol{\mu}_{\phi^{(l)}} + \boldsymbol{\Sigma}_{\phi^{(v)}} + \boldsymbol{\Sigma}_{\phi^{(l)}})$ . In the *labeled-incomplete* context, the variational lower bound on the marginal likelihood for a single data point can be given by:

$$\begin{aligned} \log p_\theta(\mathbf{x}^o, y) &\geq \int_{\mathbf{z}} \int_{\mathbf{x}^m} \log p_\theta(\mathfrak{X}, y, \mathbf{z}) dz d\mathbf{x}^m \\ &= \mathbb{E}_{q_\psi(\mathbf{x}^m|\mathbf{x}^o)}[-\mathcal{LC}(\mathfrak{X}, y) - \log q_\psi(\mathbf{x}^m|\mathbf{x}^o)] \equiv -\mathcal{LI}(\mathbf{x}^o, y). \end{aligned} \quad (13)$$

The solution to  $\mathbb{E}_{q_\psi(\mathbf{x}^m|\mathbf{x}^o)}[-\log q_\psi(\mathbf{x}^m|\mathbf{x}^o)]$  is analytical since the conditional distribution  $q_\psi(\mathbf{x}^m|\mathbf{x}^o)$  is assumed to be a Gaussian (cf. Eq. (11)). For *unlabeled-complete* data point, the variational lower bound on the marginal likelihood can be obtained by

$$\begin{aligned} \log p_\theta(\mathfrak{X}) &\geq \int_{\mathbf{z}} \int_y \log p_\theta(\mathfrak{X}, y, \mathbf{z}) dz dy \\ &= \mathbb{E}_{q_\varphi(y|\mathfrak{X})}[-\mathcal{LC}(\mathfrak{X}, y) - \log q_\varphi(y|\mathfrak{X})] \equiv -\mathcal{UC}(\mathfrak{X}). \end{aligned} \quad (14)$$

For *unlabeled-incomplete* case, the variational lower bound on the marginal likelihood can be given by:

$$\begin{aligned} \log p_\theta(\mathbf{x}^o) &\geq \int_{\mathbf{z}} \int_y \int_{\mathbf{x}^m} \log p_\theta(\mathfrak{X}, y, \mathbf{z}) dz dy d\mathbf{x}^m \\ &= \mathbb{E}_{q_\psi(\mathbf{x}^m|\mathbf{x}^o)}\{\mathbb{E}_{q_\varphi(y|\mathfrak{X})}[-\mathcal{LC}(\mathfrak{X}, y) - \log q_\varphi(y|\mathfrak{X})] \\ &\quad - q_\psi(\mathbf{x}^m|\mathbf{x}^o)\} \equiv -\mathcal{UI}(\mathbf{x}^o). \end{aligned} \quad (15)$$

Comparing to Eq. (14) we see that aside from the explicit conditional distribution for unknown label  $y$  we have added a conditional distribution  $q_\psi(\mathbf{x}^m|\mathbf{x}^o)$  for missing view  $\mathbf{x}^m$ .

The objective function for all available data points is now:

$$\begin{aligned} \mathcal{J}_{\text{SiMVAE}} &= \underbrace{\sum_{(\mathfrak{X}, y) \in S_{Ic}} \mathcal{LC}(\mathfrak{X}, y)}_{\text{labeled-complete}} + \underbrace{\sum_{(\mathbf{x}^o, y) \in S_{Ii}} \mathcal{LI}(\mathbf{x}^o, y)}_{\text{labeled-incomplete}} \\ &+ \underbrace{\sum_{\mathfrak{X} \in S_{uc}} \mathcal{UC}(\mathfrak{X})}_{\text{unlabeled-complete}} + \underbrace{\sum_{\mathbf{x}^o \in S_{ui}} \mathcal{UI}(\mathbf{x}^o)}_{\text{unlabeled-incomplete}}. \end{aligned} \quad (16)$$

Model performance can be improved by introducing explicit imputation loss and classification loss for complete data and labeled data, respectively. Therefore, the final objective function is

$$\begin{aligned} \mathcal{F}_{\text{SiMVAE}} &= \mathcal{J}_{\text{SiMVAE}} + \alpha_1 \cdot \sum_{\mathfrak{X} \in S_c} [-\log q_\psi(\mathbf{x}^m|\mathbf{x}^o)] \\ &+ \alpha_2 \cdot \sum_{(\mathfrak{X}, y) \in S_l} [-\log q_\varphi(y|\mathfrak{X})], \end{aligned} \quad (17)$$

where  $\alpha_1$  and  $\alpha_2$  are weight parameters,  $S_c = S_{Ic} \cup S_{uc}$  and  $S_l = S_{Ic} \cup S_{Ii}$ . We set  $\alpha_1 = c_1 \cdot \frac{(N_c + N_i)}{N_c}$  and  $\alpha_2 = c_2 \cdot \frac{(N_l + N_u)}{N_l}$ , where  $c_1$  and  $c_2$  are scaling constants, and  $N_c$ ,  $N_i$ ,  $N_l$  and  $N_u$  are the numbers of complete, incomplete, labeled and unlabeled data in one minibatch, respectively. Noted that the explicit classification loss (i.e., last term in Eq. (17)) allows SiMVAE to use the partially observed category information to assist the generation of  $\mathbf{x}^m$  given  $\mathbf{x}^o$ , which is more effective than the unsupervised imputation algorithms [6, 37]. Similarly, Eq. (17) can be optimized by using the stochastic backpropagation technique [15, 27].

In principle, our SiMVAE can also handle multiple missing views simultaneously. The formulas are omitted here since they can be derived straightforwardly by using multiple distinct conditional density functions  $q_\psi(\mathbf{x}^m|\mathbf{x}^o)$ .

### 3.3.4 Connections to auxiliary deep generative models.

Maaløe et al. [20] proposed auxiliary DGMs (ADGM and SDGM) by defining the inference model as  $q_\psi(\mathbf{a}|\mathbf{x}^o)q_\varphi(y|\mathbf{a}, \mathbf{x}^o)q_\phi(\mathbf{z}|\mathbf{a}, \mathbf{x}^o, y)$ , where  $\mathbf{a}$  is the auxiliary variable introduced to make the variational distribution more expressive, and  $q_\psi(\mathbf{a}|\mathbf{x}^o) = \mathcal{N}(\boldsymbol{\mu}_\psi(\mathbf{x}^o), \text{diag}(\boldsymbol{\sigma}_\psi^2(\mathbf{x}^o)))$ . If  $\mathbf{x}^m$  is a totally unobservable variable in Figures 1e and 1f, similar to SDGM, SiMVAE becomes a two-layered stochastic model. Since the generative process is conditioned on the auxiliary variable, two-layered stochastic model is more flexible than ADGM [20]. Standard ADGM and SDGM could not handle incomplete multi-view data. We endow them with this ability by forcing the inferred auxiliary variable  $\mathbf{a}$  close to  $\mathbf{x}^m$  on the set of complete data. E.g., we can obtain the objective function of SDGM+ by introducing an additional imputation loss to SDGM:

$$\mathcal{F}_{\text{SDGM+}} = \mathcal{J}_{\text{SDGM}} + \alpha_3 \cdot \sum_{\mathfrak{X} \in S_c} [-\log q_\psi(\mathbf{a}|\mathbf{x}^o)], \quad (18)$$

where  $\alpha_3$  is a regularization parameter,  $\mathfrak{X} = \{\mathbf{x}^m, \mathbf{x}^o\}$  and  $S_c$  denotes the set of complete data.  $\mathcal{J}_{\text{SDGM}}$  can be found in [20]. Intuitively, SDGM+ not only enjoys the advantages of SDGM (in terms of flexibility, convergence and performance), but also captures the relationships between views via the auxiliary inference model  $q_\psi(\mathbf{a}|\mathbf{x}^o)$ . However, SDGM+ sets a single Gaussian in the

variational distribution  $q_\phi(\mathbf{z}|\mathbf{a}, \mathbf{x}^o, \mathbf{y})$ , which may restrict its ability in multi-modality fusion.

## 4 EXPERIMENTS

We conduct experiments on two multi-modal emotion datasets to demonstrate the effectiveness of the proposed framework.

### 4.1 Datasets

**SEED:** The SEED dataset [45] contains Electroencephalogram (EEG) and eye movement (Eye) signals from 9 subjects during watching 15 movie clips, where each movie clip lasts about 4 minutes long. The EEG signals were recorded from 62 channels and the Eye signals contained information about blink, saccade fixation and so on. We used the EEG and Eye data from 9 subjects across 3 sessions, totally 27 data files. For each data file, data from watching the 1-9 movie clips were used as training set, while data from watching the 10-12 movie clips were used as validation set and the rest (13-15) were used as testing set.

**DEAP:** The DEAP dataset [16] contains EEG and peripheral physiological signals (PPS) from 32 subjects during watching 40 one-minute duration music videos. The EEG signals were recorded from 32 channels, whereas the PPS was recorded from 8 channels. The participants, using values from 1 to 9, rated each music video in terms of the levels of valence, arousal and so on. In our experiment, the valence-arousal space was divided into four quadrants according to the ratings. The threshold we used was 5, leading to four classes of data. Considering the variations of participants' ratings possibly associated with individual difference in rating scale, we discarded the samples whose ratings of arousal and valence are between 4 and 6. The dataset was randomly divided into 10-folds, where 8 folds for training, one fold for validation and the last fold for testing.

For SEED, we used the extracted differential entropy (DE) features and eye movement features (blink, saccade fixation and so on) [18]. For DEAP, following [18], we split the time series data into many one-second non-overlapping segments, where each segment is treated as an instance. Then we extracted the DE features from EEG and PPS data instances. The DE features can be calculated in four frequency bands: theta (4-8Hz), alpha (8-14Hz), beta (14-31Hz), and gamma (31-45Hz), and we used all band's features. The details of the data used in our experiments were summarized in Table 1.

dataset	#sample	#modality (#dim.)	#training	#validation	#test	#class
SEED	22734	EEG(310), Eye(33)	13473	4725	4536	3
DEAP	21042	EEG(128), PPS(32)	16834	2104	2104	4

Table 1: Properties of the data used in experiments.

### 4.2 Semi-supervised Classification with Multi-Modality Emotional Data

**4.2.1 Experimental setting.** To simulate SSL scenario, on both datasets, we randomly labeled different proportions of samples in the training set, and remained the rest samples in the training set unlabeled. For transductive SSL, we trained models on the dataset consisting of the testing data and labeled data belonging

to training set. For inductive SSL, we trained models on the entire training set consisting of the labeled and unlabeled data. For supervised learning, we trained models on the labeled data belonging to training set, and test their performance on the testing set. We compared our SMVAE with a broad range of solutions, including MAE [21], DCCA [2], DCCAE [37], AMMSS [4], AMGL [22], M2 [13] and SDGM [20]. For SMVAE, we considered multi-layer perceptrons as the type of inference and generative networks. On both datasets, we set the hidden architectures of the inference and generative networks for each view as '100-50-30' and '30-50-100', respectively, where '30' is the dimension of the latent variables. We used the Adam optimizer [12] with a learning rate  $\eta = 3 \times 10^{-4}$  in training. The scaling constant  $c$  was selected from {0.1, 0.5, 1}. For MAE, DCCA and DCCAE, we considered the same setups (network structure, learning rate, etc.) as our SMVAE. Furthermore, we used support vector machines (SVM) and transductive SVM (TSVM) for supervised learning and transductive SSL, respectively. For AMGL, M2 and SDGM we used their default settings, and we evaluated M2's performance on each modality and the concatenation of all modalities, respectively.

SEED data	Algorithms	1% labeled	2% labeled	3% labeled
Supervised learning	MAE+SVM [21]	.814±.031	.896±.024	.925±.024
	DCCA+SVM [2]	.809±.035	.891±.035	.923±.028
	DCCAE+SVM [37]	.819±.036	.893±.034	.923±.027
Transductive SSL	AMMSS [4]	.731±.055	.839±.036	.912±.018
	AMGL [22]	.711±.047	.817±.023	.886±.028
	MAE+TSVM [21]	.818±.035	.910±.025	.931±.026
	DCCA+TSVM [2]	.811±.031	.903±.024	.928±.021
	DCCAE+TSVM [37]	.823±.040	.907±.027	.929±.023
	SMVAE	<b>.861±.037</b>	<b>.931±.020</b>	<b>.960±.021</b>
Inductive SSL	M2 (Eye) [13]	.753±.024	.849±.055	.899±.049
	M2 (EEG) [13]	.768±.041	.861±.040	.919±.026
	M2 (Concat.) [13]	.803±.035	.876±.043	.926±.044
	SDGM (Concat.) [20]	.819±.034	.893±.042	.932±.041
	SMVAE	<b>.880±.033</b>	<b>.955±.020</b>	<b>.968±.015</b>

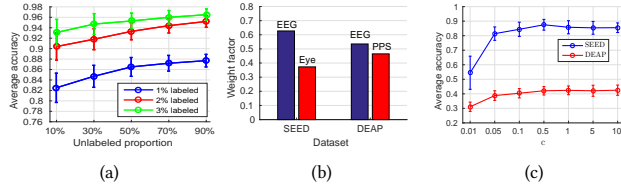
DEAP data	Algorithms	1% labeled	2% labeled	3% labeled
Supervised learning	MAE+SVM [21]	.353±.027	.387±.014	.411±.016
	DCCA+SVM [2]	.359±.016	.400±.014	.416±.018
	DCCAE+SVM [37]	.361±.023	.403±.017	.419±.013
Transductive SSL	AMMSS [4]	.303±.029	.353±.024	.386±.014
	AMGL [22]	.291±.027	.341±.021	.367±.019
	MAE+TSVM [21]	.376±.025	.403±.031	.417±.026
	DCCA+TSVM [2]	.379±.021	.408±.024	.421±.017
	DCCAE+TSVM [37]	.384±.022	.412±.027	.425±.021
	SMVAE	<b>.424±.020</b>	<b>.441±.013</b>	<b>.456±.013</b>
Inductive SSL	M2 (PPS) [13]	.366±.024	.389±.048	.402±.034
	M2 (EEG) [13]	.374±.019	.397±.013	.407±.016
	M2 (Concat.) [13]	.383±.019	.404±.016	.416±.015
	SDGM (Concat.) [20]	.389±.019	.411±.017	.423±.015
	SMVAE	<b>.421±.017</b>	<b>.439±.015</b>	<b>.451±.013</b>

Table 2: Comparison of classification accuracies with different proportions of labeled training samples.

**4.2.2 Classification accuracy with very few labels.** Table 2 presents the classification accuracies of all methods on SEED and DEAP datasets. The proportions of labeled samples in the training set vary from 1% to 3%. Results (mean±std) were averaged over 20 independent runs. Several observations can be drawn as follows. First, the average accuracy of SMVAE significantly surpasses the baselines in all cases. Second, by examining SMVAE against supervised learning approaches trained on very limited labeled data, we can find that SMVAE always outperforms them. This encouraging result shows that SMVAE can effectively leverage the useful information from unlabeled data. Third, multi-view semi-supervised algorithms

AMSS and AMGL perform worst in all cases. We attribute this to the fact that graph-based shallow models AMSS and AMGL cannot extract the deep features from the original data. Fourth, the performances of three TSVM-based semi-supervised methods are moderate. Finally, compared with the single-view methods M2 and SDGM, our multi-view method is more effective in integrating multiple modalities.

**4.2.3 Flexibility and stability.** The proportion of unlabeled samples in the training set will affect the performance of semi-supervised models. Figure 2a shows the changes of inductive SiMVAE's average accuracy on SEED with different proportions of unlabeled samples in the training set. We can observe that the unlabeled samples can effectively boost the classification accuracy of SiMVAE. Instead of treating each modality equally, SiMVAE can weight each modality and perform classification simultaneously. Figure 2b shows the learned weight factors by inductive SiMVAE on both datasets (1% labeled). From it, we can observe that EEG modality has the highest weight on both datasets, which is consistent with single modality's performance of M2 shown in Table 2 and the results in previous work [18]. The scaling constant  $c$  controls the weight of discriminative learning in SiMVAE. Figure 2c shows the performance of inductive SiMVAE with different  $c$  values (1% labeled). From it, we can find that the scaling constant  $c$  can be chosen from  $\{0.1, 0.5, 1\}$ , where SiMVAE achieves good results.



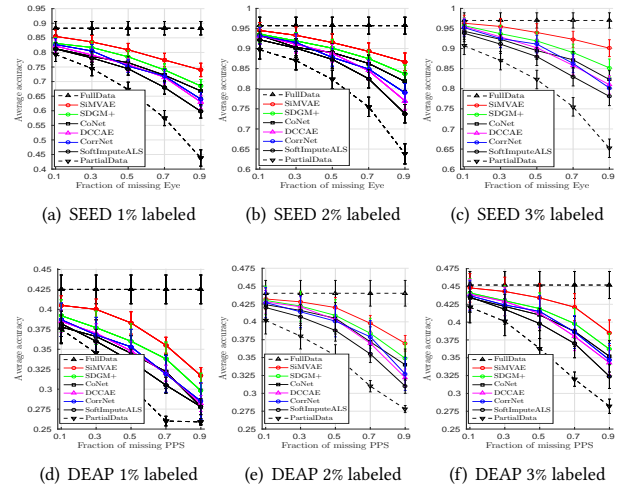
**Figure 2: For Inductive SiMVAE: (a) performance with different proportions of unlabeled training samples on SEED dataset, (b) learned weight factors, (c) the impact of scaling constant  $c$ .**

### 4.3 Semi-supervised Learning with Incomplete Multi-Modality Data

**4.3.1 Experimental setting.** To simulate the incomplete data setting, we randomly selected a fraction of instances (from both labeled and unlabeled training data) to be unpaired examples, i.e., they are described by only one modality, and the remaining ones appear in both modalities. We varied the fraction of missing data from 10% to 90% with an interval of 20%, while no missing data in validation and testing sets. In our experiment, we assumed the Eye modality of SEED and the PPS modality of DEAP are incomplete.

There are two main solutions for semi-supervised classification of incomplete multi-view data. One way is to complete the missing view firstly in an unsupervised way, and then conduct semi-supervised classification. Another way is to integrate missing view imputation and semi-supervised classification into an end-to-end learning framework. We compared our (inductive) SiMVAE

algorithm with these two ways. Specifically, we compared SiMVAE with SoftImputeALS [8], DCCAE [37], CorrNet [6], CoNet [25] and SDGM+ (a variant of SDGM [20], cf. Section 3.3.4). For SoftImputeALS, DCCAE and CorrNet, we first estimated the missing modalities by using the authors' implementation, and then conducted semi-supervised classification by using our (inductive) SiMVAE algorithm. For CoNet and SDGM+, we conducted missing modality imputation and semi-supervised classification simultaneously based on our own implementations. Additionally, we also compared SiMVAE with the following two baselines: 1) SiMVAE with complete data (FullData, i.e., no missing modality for any training instances), which can be regarded as an upper bound of SiMVAE; 2) SiMVAE with only paired data (PartialData, i.e., we simply discard those incomplete samples in training process), which can be regarded as a lower bound of SiMVAE. These two bounds define the potential range of SiMVAE's performance. For SiMVAE, both  $c_1$  and  $c_2$  were selected from  $\{0.1, 0.5, 1\}$ . For SDGM+, we selected the regularization parameter  $\alpha_3$  from  $\{1e-3, 1e-2, \dots, 1e3\}$ .

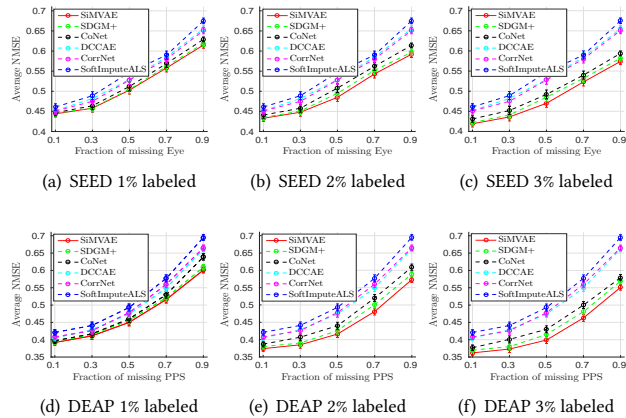


**Figure 3: Comparison of recognition accuracies with different fractions of missing data and labeled data.**

**4.3.2 Semi-supervised classification.** The performance of our SiMVAE and the compared methods was shown in Figure 3, where each point on every curve is an average over 20 independent trials. From Figure 3, it is seen that SiMVAE consistently outperforms the compared methods. Compared with the two-stage methods (SoftImputeALS, DCCAE and CorrNet), the advantage of SiMVAE is significant, especially when there are sufficient labeled data (3%). This is because SiMVAE can make good use of the available category information to generate more informative modalities, which in turn will improve classification performance. Whereas the two-stage methods couldn't obtain the global optimal results. Also, SiMVAE shows obvious advantage over the semi-supervised methods CoNet and SDGM+. This may be because CoNet and SDGM+ are not designed to integrate multiple modalities. Moreover, SiMVAE has been successful even when a high percentage of samples

are incomplete. Specifically, SiMVAE with even about 50% incomplete samples achieves comparable results to the fully complete case (FullData). With fractions lower than that, we observe that SiMVAE roughly reached FullData’s performance, especially when the labeled data are sufficient. Finally, SiMVAE’s performance is more closer to FullData than to PartialData, which indicates the effectiveness of SiMVAE in learning from incomplete data.

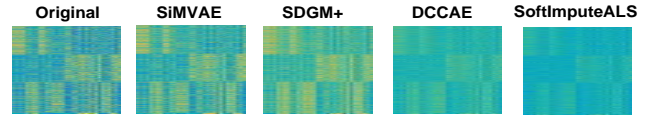
**4.3.3 Missing modality imputation.** Since the quality of recovered missing modalities directly affects the classification results, we also evaluated the performance of missing modality imputation for all methods. For SiMVAE and SDGM+, we obtained the single imputation of  $\mathbf{x}^m$  by evaluating the conditional mean ( $\mathbf{x}^m = \mathbb{E}[q_{\psi}(\mathbf{x}^m | \mathbf{x}^o)]$ ). We used the Normalized Mean Squared Error (NMSE) to measure the relative distance between the original and the recovered modalities.  $NMSE = \frac{\|\mathbf{X} - \hat{\mathbf{X}}\|_F}{\|\mathbf{X}\|_F}$ , where  $\mathbf{X}$  and  $\hat{\mathbf{X}}$  are the original and the recovered data matrices, respectively.  $\|\cdot\|_F$  demotes the Frobenius norm. Figure 4 shows the experimental results.



**Figure 4: Comparison of imputation errors with different fractions of missing data and labeled data.**

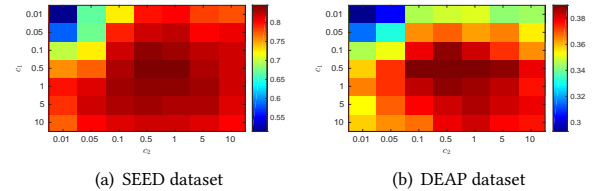
From Figure 4, it can be seen that as the fraction of missing data increases, the relative distance between the original modalities and the recovered modalities increases. Further, the semi-supervised imputation methods (SiMVAE, CoNet and SDGM+) consistently outperforms the unsupervised imputation methods (SoftImputeALS, DCCAE and CorrNet), and increasing the number of labeled training data improves the imputation performance of semi-supervised methods. This demonstrates that the category information plays an important role in missing modality imputation. SoftImputeALS shows the worst performance, which verifies that matrix completion method is not suitable for missing modality imputation. CoNet and SDGM+ obtain comparable imputation errors to SiMVAE. This indicates that their moderate classification performance in Figure 3 may be caused by their inability in modality fusion. Except for SiMVAE and SDGM+, other methods ignore the uncertainty of the missing view, which also limits their imputation performance. To compare the imputation performance more intuitively, we visualize the original and recovered data matrices in Figure 5 (on SEED, 3%

labeled and 10% missing Eye). From it, we see that SiMVAE recovered more individual characteristics of the original data matrix than other methods.



**Figure 5: Visualization of the original and the recovered data matrices on SEED dataset. Each row of each panel is an instance of the missing modality.**

**4.3.4 Sensitivity analysis.** Figure 6 shows the classification accuracies of inductive SiMVAE with different scaling constants  $c_1$  and  $c_2$  on both datasets (1% labeled and 10% missing data). From it, we can find that SiMVAE is not very sensitive to the values of  $c_1$  and  $c_2$ . We choose the best  $c_1$  and  $c_2$  from  $\{0.1, 0.5, 1\}$  in the experiments.



**Figure 6: The impact of scaling constants  $c_1$  and  $c_2$ .**

## 5 CONCLUSION

We have proposed a novel semi-supervised multi-view deep generative framework for multi-modal emotion recognition with incomplete data. Under our framework, each modality of the emotional data is treated as one view, and the importance of each modality is inferred automatically by learning a non-uniformly weighted Gaussian mixture posterior approximation for the shared latent variable. The labeled-data-scarcity problem is naturally addressed within our framework through casting the semi-supervised classification problem as a specialized missing data imputation task. The incomplete-data problem is elegantly circumvented by treating the missing views as latent variables and integrating them out. Compared with previous emotion recognition methods, our method is more robust and flexible. Experimental results confirmed the superiorities of our framework over many state-of-the-art competitors.

## ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 91520202, 61602449), Beijing Municipal Science&Technology Commission (Z181100008918010), Youth Innovation Promotion Association CAS and Strategic Priority Research Program of CAS.



## REFERENCES

- [1] Massih Reza Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from Multiple Partially Observed Views – an Application to Multilingual Text Categorization. *NIPS* (2009), 28–36.
- [2] Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*. 1247–1255.
- [3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. 2016. Importance Weighted Autoencoders. In *ICLR*.
- [4] Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. 2013. Heterogeneous Image Features Integration via Multi-modal Semi-supervised Learning Model. In *ICCV*. 1737–1744.
- [5] Rafael A Calvo and Sidney D’Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1 (2010), 18–37.
- [6] Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, and Balaraman Ravindran. 2016. Correlational neural networks. *Neural computation* 28, 2 (2016), 257–285.
- [7] Samuel Gershman, Matt Hoffman, and David Blei. 2012. Nonparametric variational inference. In *ICML*.
- [8] Trevor Hastie, Rahul Mazumder, Reza Zadeh, and Reza Zadeh. 2015. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research* 16, 1 (2015), 3367–3402.
- [9] Elad Hazan, Roi Livni, and Yishay Mansour. 2015. Classification with Low Rank and Missing Data. In *ICML*. 257–266.
- [10] Xiaowei Jia, Kang Li, Xiaoyi Li, and Aidong Zhang. 2014. A novel semi-supervised deep learning framework for affective state recognition on EEG signals. In *International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 30–37.
- [11] Raghunandan H. Keshavan, Sewoong Oh, and Andrea Montanari. 2009. Matrix completion from a few entries. *IEEE Transactions on Information Theory* 56, 6 (2009), 2980–2998.
- [12] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [13] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *NIPS*. 3581–3589.
- [14] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improving Variational Inference with Inverse Autoregressive Flow. In *NIPS*. 4743–4751.
- [15] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.
- [16] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.
- [17] Wei Liu, Wei Long Zheng, and Bao Liang Lu. 2016. Emotion Recognition Using Multimodal Deep Learning. In *International Conference on Neural Information Processing*. 521–529.
- [18] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining Eye Movements and EEG to Enhance Emotion Recognition. In *IJCAI*. 1170–1176.
- [19] Tran Luan, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *CVPR*. 4971–4980.
- [20] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. 2016. Auxiliary deep generative models. In *ICML*. 1445–1453.
- [21] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*. 689–696.
- [22] Feiping Nie, Jing Li, Xuelong Li, et al. 2016. Parameter-Free Auto-Weighted Multiple Graph Learning: A Framework for Multiview Clustering and Semi-Supervised Classification. In *IJCAI*. 1881–1887.
- [23] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. 2015. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia* 17, 11 (2015), 2008–2020.
- [24] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.
- [25] Brian Quanz and Jun Huan. 2012. CoNet: feature generation for multi-view semi-supervised learning with partially observed views. In *CIKM*. 1273–1282.
- [26] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *Applications of Computer Vision*. 1–9.
- [27] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *NIPS*. 1278–1286.
- [28] Martin Schels, Markus Kächele, Michael Glodek, David Hrabal, Steffen Walter, and Friedhelm Schwenker. 2014. Using unlabeled data to improve classification of emotional states in human computer interaction. *Journal on Multimodal User Interfaces* 8, 1 (2014), 5–16.
- [29] Iulian V Serban, II Ororbia, G Alexander, Joelle Pineau, and Aaron Courville. 2016. Multi-modal Variational Encoder-Decoders. *arXiv preprint arXiv:1612.00377* (2016).
- [30] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko Shin Chen, Jin Lu, and Jinbo Bi. 2017. VIGAN: Missing view imputation with generative adversarial networks. In *IEEE International Conference on Big Data*. 766–775.
- [31] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2016. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 7, 1 (2016), 17–28.
- [32] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. 2016. Ladder variational autoencoders. In *NIPS*. 3738–3746.
- [33] Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research* 15 (2014), 2949–2980.
- [34] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.
- [35] Johannes Wagner, Elisabeth Andre, Florian Ringenfelder, and Jonghwa Kim. 2011. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing* 2, 4 (2011), 206–218.
- [36] C. Wang, X. Liao, L. Carin, and D. B. Dunson. 2010. Classification with Incomplete Data Using Dirichlet Process Priors. *Journal of Machine Learning Research* 11, 18 (2010), 3269.
- [37] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A Bilmes. 2015. On Deep Multi-View Representation Learning. In *ICML*. 1083–1092.
- [38] Weiran Wang, Xinchun Yan, Honglak Lee, and Karen Livescu. 2016. Deep Variational Canonical Correlation Analysis. *arXiv: 1610.03454* (2016).
- [39] D Williams, X. Liao, Y. Xue, L. Carin, and B Krishnapuram. 2007. On classification with incomplete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 3 (2007), 427.
- [40] C. Xu, D. Tao, and C. Xu. 2015. Multi-view Learning with Incomplete Views. *IEEE Transactions on Image Processing* 24, 12 (2015), 5812–5825.
- [41] Shipeng Yu, Balaji Krishnapuram, Rómber Rosales, and R. Bharat Rao. 2011. Bayesian Co-Training. *Journal of Machine Learning Research* 12, 3 (2011), 2649–2680.
- [42] Lei Zhang, Yao Zhao, Zhenfeng Zhu, Dinggang Shen, and Shuiwang Ji. 2018. Multi-View Missing Data Completion. *IEEE Transactions on Knowledge and Data Engineering* 30, 7 (2018), 1296–1309.
- [43] Zixing Zhang, Fabien Ringeval, Bin Dong, Eduardo Coutinho, Erik Marchi, and Björn Schüller. 2016. Enhanced semi-supervised learning for multimodal emotion recognition. In *ICASSP*. IEEE, 5185–5189.
- [44] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Transactions on Cybernetics* (2018), 1–13.
- [45] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (2015), 162–175.

## Supplementary Material

In this document, we provide additional materials to supplement our main submission. In Section A, we provide further details on how we derive a second lower bound to variational lower bound. In Section B, we show Monte-Carlo estimators used to compute the gradients of the objective function.

### SECTION A

The Shannon entropy  $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[-\log q_\phi(\mathbf{z}|\mathbf{x}, y)]$  is hard to compute analytically. In general, there is no closed-form expression for the entropy of a Mixture of Gaussians (MoG). Here we lower bound the entropy of MoG using Jensen's inequality:

$$\begin{aligned}
& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[-\log q_\phi(\mathbf{z}|\mathbf{x}, y)] \\
&= - \int q_\phi(\mathbf{z}|\mathbf{x}, y) \log q_\phi(\mathbf{z}|\mathbf{x}, y) d\mathbf{z} \\
&= - \sum_{v=1}^2 \lambda^{(v)} \cdot \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}}, \boldsymbol{\Sigma}_{\phi^{(v)}}) \log \sum_{l=1}^2 \lambda^{(l)} \\
&\quad \cdot \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(l)}}, \boldsymbol{\Sigma}_{\phi^{(l)}}) d\mathbf{z} \\
&\geq - \sum_{v=1}^2 \lambda^{(v)} \cdot \log \sum_{l=1}^2 \lambda^{(l)} \cdot \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}}, \boldsymbol{\Sigma}_{\phi^{(v)}}) \\
&\quad \cdot \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(l)}}, \boldsymbol{\Sigma}_{\phi^{(l)}}) d\mathbf{z} \\
&= - \sum_{v=1}^2 \lambda^{(v)} \cdot \log \sum_{l=1}^2 \lambda^{(l)} \cdot \mathcal{N}(\boldsymbol{\mu}_{\phi^{(v)}}|\boldsymbol{\mu}_{\phi^{(l)}}, \boldsymbol{\Sigma}_{\phi^{(v)}} + \boldsymbol{\Sigma}_{\phi^{(l)}}) \\
&= - \sum_{v=1}^2 \lambda^{(v)} \cdot \log \left( \sum_{l=1}^2 \lambda^{(l)} \cdot \omega_{v,l} \right),
\end{aligned}$$

where we have used the fact that the convolution of two Gaussians is another Gaussian, and  $\omega_{v,l} = \mathcal{N}(\boldsymbol{\mu}_{\phi^{(v)}}|\boldsymbol{\mu}_{\phi^{(l)}}, \boldsymbol{\Sigma}_{\phi^{(v)}} + \boldsymbol{\Sigma}_{\phi^{(l)}})$ .

### SECTION B

$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})]$  can be rewritten, using the location-scale transformation for the Gaussian distribution, as:

$$\begin{aligned}
& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})] \\
&= \sum_{l=1}^2 \lambda^{(l)} \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0}, \mathbf{I})} \left[ \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \boldsymbol{\mu}_{\phi^{(l)}} + \mathbf{R}_{\phi^{(l)}} \boldsymbol{\epsilon}^{(l)}) \right],
\end{aligned}$$

where  $\mathbf{R}_{\phi^{(l)}} \mathbf{R}_{\phi^{(l)}}^\top = \boldsymbol{\Sigma}_{\phi^{(l)}}$  and  $l \in \{1, 2\}$ . While the expectations on the right hand side of the above equation still cannot be solved analytically, their gradients w.r.t.  $\theta^{(v)}$ ,  $\phi^{(l)}$  and  $\lambda^{(l)}$  can be efficiently estimated using the following Monte-Carlo estimators

$$\begin{aligned}
& \frac{\partial}{\partial \theta^{(v)}} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})] \\
&= \sum_{l=1}^2 \lambda^{(l)} \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0}, \mathbf{I})} \left[ \frac{\partial}{\partial \theta^{(v)}} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}^{(l)}) \right] \\
&\approx \frac{\lambda^{(l)}}{T} \sum_{t=1}^T \sum_{l=1}^2 \frac{\partial}{\partial \theta^{(v)}} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}^{(l,t)}),
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial}{\partial \phi^{(l)}} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})] \\
&= \lambda^{(l)} \frac{\partial}{\partial \phi^{(l)}} \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0}, \mathbf{I})} \left[ \frac{\partial}{\partial \mathbf{z}^{(l)}} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}^{(l)}) \right. \\
&\quad \left. \cdot \left( \frac{\partial \boldsymbol{\mu}_{\phi^{(l)}}}{\partial \phi^{(l)}} + \frac{\partial \mathbf{R}_{\phi^{(l)}}}{\partial \phi^{(l)}} \boldsymbol{\epsilon}^{(l)} \right) \right] \\
&\approx \frac{\lambda^{(l)}}{T} \sum_{t=1}^T \frac{\partial}{\partial \mathbf{z}^{(l,t)}} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}^{(l,t)}) \\
&\quad \cdot \left( \frac{\partial \boldsymbol{\mu}_{\phi^{(l)}}}{\partial \phi^{(l)}} + \frac{\partial \mathbf{R}_{\phi^{(l)}}}{\partial \phi^{(l)}} \boldsymbol{\epsilon}^{(l,t)} \right), \\
& \frac{\partial}{\partial \lambda^{(l)}} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z})] \\
&= \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0}, \mathbf{I})} [\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}^{(l)})] \\
&\approx \frac{1}{T} \sum_{t=1}^T \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y, \mathbf{z}^{(l,t)}),
\end{aligned}$$

where  $\mathbf{z}^{(l)}$  is evaluated at  $\mathbf{z}^{(l)} = \boldsymbol{\mu}_{\phi^{(l)}} + \mathbf{R}_{\phi^{(l)}} \boldsymbol{\epsilon}^{(l)}$  and  $\mathbf{z}^{(l,t)} = \boldsymbol{\mu}_{\phi^{(l)}} + \mathbf{R}_{\phi^{(l)}} \boldsymbol{\epsilon}^{(l,t)}$  with  $\boldsymbol{\epsilon}^{(l,t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . In practice, it suffices to use a small  $T$  (e.g.  $T = 1$ ) and then estimate the gradient using mini-batches of data points. Though the above Monte-Carlo estimators could have large variances if a small  $T$  is used, the experimental results show that it suffices to obtain good performance. The same observation can be found in previous works [20, 32]. Furthermore, we use the same random numbers  $\boldsymbol{\epsilon}^{(l,t)}$  for all estimators to have lower variances. The gradient w.r.t.  $\phi$  is omitted here, since it can be derived straightforwardly by using traditional reparameterization trick [13].